The Use and Usefulness of p-Values in Political Science: Introduction

Bischof, Daniel ; Van Der Velden, Mariken

DOI: https://doi.org/10.1111/spsr.12376

The Use and Usefullness of *P*-values in Political Science: Intro

Daniel Bischof[†]
Mariken van der Velden.

2nd September 2021

P-values are the most frequently employed metric to assess the significance of statistical findings in the social sciences. Since the earliest years of their usage the meaning and usefulness of P-values were topics of heated discussion (Fisher 1935; Berkson 1942). Lately the reproduction/replication crisis resuscitated this debate (Gelman 2018; Benjamin et al. 2018; Lakens et al. 2018; McShane et al. 2017; Trafimow and Marks 2015; Nuzzo 2014). Meanwhile, the skepticism has not stopped at the gates of political science. Most prominently the journal "Political Analysis" banned P-values "in regression tables or elsewhere" after the new editor took over the board of editors in 2017 (Gill 2018: 1).[1] Also political scientists contributed to a swelling debate suggesting to lower the threshold for P-values to 0.005 (Benjamin et al. 2018; Esarey 2017).

This special issue seeks to contribute to the debate on P-values by summarizing the main arguments of it, providing an encompassing discussion of P-values – also from an epistemological perspective – as well as advice for the discipline about the Do's and Don'ts for P-values. In February 2018 the Department of Political Science at the University of Zurich invited several political methodologists to discuss the matter in a series of public lectures. The present contribution summarizes these public lectures but also goes beyond them by presenting the arguments two other distinguished colleagues.

Our introductory piece summarizes the discussion around P-values in the discipline and situates this into the larger debate on the replication crisis; Justin Esarey discusses what null

[†]Department of Political Science, University of Zurich (CH); corresponding author: bischof@ipz.uzh.ch.
[1]Only to back-pedal later on and deciding to allow P-values in specific cases.

hypothesis testing means for social sciences in general as well as for the P-value debate; Susumu Shikano provides a detailed insight into how Bayesian inference fails and merits of Bayesian inference in addressing the P-value crisis; Marco Steenbergen takes an epistemological view on to P or not to P; Vera Tröger talks about the use of P-values from an econometric perspective; Simon Hug's contribution finally synthesizes our debate.

So, what are P-values and why do we as researchers care so much about them? P-values date back to the 18^{th} century where Pierre Simon-Laplace came-up with first ad hoc definitions of P-values. Later on in the 20^{th} century Pearson formalized them in his χ^2 test and then Fisher popularized them by also proposing the threshold of 0.05 for statistical significance. The first references from Fisher on the threshold of 0.05 stems from his well-known "lady testing tea" experiment. Muriel Bristol, a phycologist and enthusiastic tea drinker, claimed to be able to differentiate tea which was poured on milk from milk which was poured on tea – of course keeping the amount of tea and milk constant.

Fisher and his friend William Roach decided to test Bristol's tea tasting skills with a simple experiment: Muriel Bristol was provided eight cups of tea (four prepared by first adding milk; four prepared by first adding tea). Bristol then was asked to name the four cups prepared by her be-liked method. Thus, the null hypothesis was that Burial did not have the ability to distinguish the preparation of tea. Given n=8 cups and k=4 chosen cups the experiment results in 70 possible combinations.[2] In order to reject the null hypothesis Fisher suggested that Bristol needed to get four out of four cups right. The combination of four correctly classified cups has a chance to occur in one out of 70 combinations. Bristol eventually got all eight cups correct.

Fisher discusses the threshold of 5 % in close relation to the lady testing tea experiment. As outlined above Bristol's performance had a chance to occur in only 1.4 per cent, while if she had missed only a single cup the chances to observe such a performance would have increased drastically to 24.3 per cent.[3] In the latter case Fisher believed the likelihood of observing

[2] $$\frac{8!}{4!(8-4)!} = 70$$

[3] $$\frac{16+1}{70} = 0.243$$

such a performance just by chance was too high. Future research built on Fisher's reasoning and eventually stopped discussing the reasons for the 5 % threshold entirely. As this example illustrates, what P-values then really tell us is how likely our data are, assuming that our H_0 (*Bristol not having the skills to tell the difference between the two tea preparation methods*) is true (Wasserstein and Lazar 2016). A standard for empirical testing was born and until today this standard guides social scientists' behavior, evaluations of research and most prominently publication standards.

But why have researchers recently 'seen the light' and payed attention to the shortcomings of P-values? In 2011, the renowned social-psychologist Diederik Stapel was found guilty of fraudulent research practices.[4] As it turned out Stapel had faked his entire data collection. He simply answered to his questionnaires himself and thereby created the data he and his research team then analyzed. His research fraud sparked a larger debate within psychology: To what extend was there a culture of "sloppy" science, in which some scientists did not understand the essentials of statistics, reviewers for journals encouraged researchers to leave unwelcome data out of their papers, and even the most prestigious journals printed results that were obviously too good to be true?[5] The Open Science Collaboration (2015) replicated 100 studies published in psychology journals. Using high-powered designs, they found that their mean effect size was approximately half of the size of the original articles. Moreover, while 97% of the original studies had demonstrated significant results ($p < 0.05$), only 47% of the replicated studies had significant results – indicating that 53% of the studies could not be replicated. This is not only a problem of psychology, where the norm is to publish based on experimental studies. In economics, also half of the studies could not be replicated (Chang and Li 2015). Chang and Li (2015) replicate 67 original articles published in 13 well-regarded macro-economics and general interest economic journals and demonstrate that replication issues are not tied to using experimental data, but equally apply to studies using publicly available data sets. In political science the debate caught fire with the Mike LaCour case. LaCour did not only follow "sloppy" research practices but committed fraud by inventing data he never had collected in the first place (Broockman, Kalla, and Aronow 2015).

[4] http://www.apa.org/science/about/psa/2011/12/diederik-stapel.aspx
[5] http://www.sciencemag.org/news/2012/11/final-report-stapel-affair-points-bigger-problems-social-psyc

These happenings suggest that practices of 'P-hacking' are more likely to occur in an environment focusing so much on the question of whether $P < 0.05$. They are then amplified by the human tendencies of *apophenia* - seeing patterns in random data - and of *conformation bias* - focusing on evidence that is in line with our (favored) explanation. These human tendencies are likely to affect how we walk through the 'garden of forking paths' when conducting analysis (Gelman and Loken 2013) and how we interact with the question of 'researcher degrees of freedom' (Simmons, Nelson, and Simonsohn 2011). And in many instances making the threshold is just one tiny step away – e.g. by adding/dropping a control, an interaction term or dropping some unfavorable outliers.

Thus, from our point of view not only the practices of how we engage and interpret P-values need to change, but eventually the environment under which we conduct research needs to adapt as well. Lowering the threshold for significance is unlikely to achieve this goal (Benjamin et al. 2018). A design-based derivation of the threshold might be better-equipped to achieve this goal (Lakens et al. 2018), but similar to the issue of the 'garden of forking paths' leaves researchers potentially with too many 'degrees of freedom'. Proposals calling for a purely Bayesian approach to questions of significance tend to ignore that eventually we will run in very similar questions and issues irrespective if we choose a Bayesian or Frequentist perspective.

Instead, we understand the replication/reproduction crisis as a symptom for a larger, systematic problem in the Social Sciences. This problem speaks to all aspects of what we are as a profession. It speaks to: how we teach empirical research practices, how we engage with changing practices in data sciences and how we question our own past and present behavior as scientists. But most importantly it suggests that no matter how we engage with P-values in the future as a profession, proposals for change need to take into account how much publication pressures, publication and review practices will affect proposed reforms to engage with P.

Daniel adds a synthesis of the proposals discuss in the special issue

References

Benjamin, Daniel J, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer et al. 2018. "Redefine statistical significance." *Nature Human Behaviour* 2 (1): 6.

Berkson, Joseph. 1942. "Tests of Significance Considered as Evidence." *Journal of the American Statistical Association* 37 (219): 325–335.

Broockman, David, Joshua Kalla, and Peter M Aronow. 2015. "Irregularities in LaCour (2014) Timeline of Disclosure." https://stanford.edu/$\sim$dbroock/broockman_kalla_aronow_lg_irregularities.pdf.

Chang, Andrew, and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'." Finance and Economics Discussion Series Divisions, url: https://www.federalreserve.gov/econresdata/feds/2015/files/2015083pap.pdf.

Esarey, Justin. 2017. "Lowering the Threshold of Statis- tical Significance to p < 0.005 to Encourage Enriched Theories of Politics." *The Political Methodologist* 24 (2): 13–19.

Fisher, Ronald Aylmer. 1935. *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Gelman, Andrew. 2018. "Ethics in statistical practice and communication." *Significance* 138 (83): 40–43.

Gelman, Andrew, and Eric Loken. 2013. "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis." *Downloaded January* pp. 1–17.

Gill, Jeff. 2018. "Comments from the new Editor." *Political Analysis* 26 (1): 1–2.

Lakens, D, JA Grange, F Adolfi, C Albers, F Anvari, M Apps, S Argamon, T Baguley, R Becker, S Benning et al. 2018. "Justify your alpha." *Nature Human Behavior* .

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert, and Jennifer L Tackett. 2017. "Abandon statistical significance." *arXiv preprint arXiv:1709.07588* .

Nuzzo, Regina. 2014. "Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume." *Nature* 506 (7487): 150–152.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22 (11): 1359–1366.

Trafimow, David, and Michael Marks. 2015. "Editorial." *Basic and Applied Social Psychology* 37 (1): 1–2.

References

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. “The ASA’s Statement on p -Values: Context, Process, and Purpose.” *The American Statistician* 70 (2): 129–133.